University of Calabria

Department of Mathematics and Computer Science

Master's Degree in Computer Science

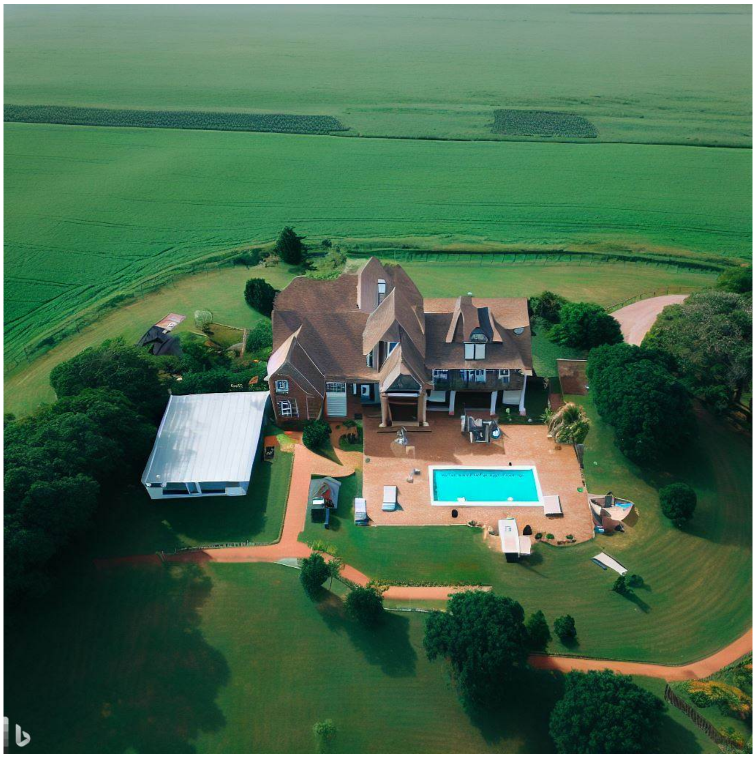
Machine Learning

Machine Learning - House Prices Analysis

Daniele Avolio Alessandro Fazio Merem Hassem Indiris Michele Vitale

Lorenzo Piro

A.Y. 2022/2023



1 Introduction

To develop this **Machine Learning** project we are going to use the CRISP-DM methodology, which is a well-known and widely used methodology for data mining projects. It is an iterative process that is composed of six phases.

In particual, the phases are:

1. **Business Understanding**: in this phase we will try to understand the problem and the objectives of the project. We will also try to understand the data that we have at our disposal and how we can use it to solve the problem.

2. **Data Understanding**: in this phase we will try to understand the data that we have at our disposal. We will try to understand the meaning of the data and how we can use it to solve the problem.

3. **Data Preparation**: in this phase we will try to prepare the data for the next phases. We will try to clean the data and to transform it in a way that will be useful for the next phases.

4. **Modeling**: in this phase we will try to build a model that will be able to solve the problem. We will try to find the best model for our problem.

5. **Evaluation**: in this phase we will try to evaluate the model that we have built. We will try to understand if the model is good enough to solve the problem.

6. **Deployment**: in this phase we will try to deploy the model that we have built. We will try to understand how we can use the model to solve the problem.

2 Business understanding

2.1 Background

Our project is based using the dataset of the Kaggle competition House Prices: Advanced Regression Techniques.

The goal of the competition is to predict the final price of each home based on 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa. The dataset is composed by 1460 rows and 81 columns, where the last column is the target variable, the Sale Price.

A small list of the variables is the following:

- **LotArea**: Lot size in square feet

- **OverallQual**: Overall material and finish quality

- **OverallCond**: Overall condition rating

- **YearBuilt**: Original construction date

- **YearRemodAdd**: Remodel date

- **RoofStyle**: Type of roof

- **Exterior1st**: Exterior covering on house

- **Exterior2nd**: Exterior covering on house (if more than one material)
- **MasVnrType**: Masonry veneer type
- **MasVnrArea**: Masonry veneer area in square feet
- ...

2.2 Business objectives

In this part we are going to analyze the business objective of the project. Our goal is different from the one of the competition, in fact we are requested to **convert the Sale Price variable into 3 ranges**:

1. **LOW**: From 0 to 150000
2. **MEDIUM**: From 150000 to 300000
3. **HIGH**: From 300000 and beyond

So, our goal is to predict a categorical variable with 3 possible values, instead of a continuous variable. This is a very important difference, because we are not interested in the exact value of the Sale Price, but only in the range in which it falls.

2.3 Business success criteria

The success criteria of the project is to obtain a model that is able to predict the Sale Price range with a good accuracy. In particular, we want to create a model that is able to predict the Sale Price range with an accuracy of at least SOMETHING.

2.4 Assessment of the situation

There is a very important aspect that we have to consider: there are a lot of variables in the dataset that contains a values of **NA**, that could leat to thinking that the value is missing.

Actually, this is not always true. In particular, the description of the dataset contains information about the meaning of the **NA** value for each variable. For example, the **NA** value for the **PoolQC** variable means that the house doesn't have a pool, so the value is not missing, but it is a value that has a meaning.

There are more variables that have a similar meaning for the **NA** value, so we have to be careful when we are going to handle the missing values. For this project, we are going to assume that the **NA** value is not missing but is just a value that has a meaning. If we need to drop the column, will be specified in the particular section

2.5 Inventory of resources

For this project, we are going to use the following resources:

- **Python 3.10.6**: The programming language used for the project
- **Jupyter Notebook**: The IDE used for the project
- **Pandas**: The library used for the data manipulation

- **Numpy**: The library used for the data manipulation
- **Matplotlib**: The library used for the data visualization
- **Seaborn**: The library used for the data visualization
- **Scikit-learn**: The library used for the machine learning

2.6 Project Plan

The development of the project will be divided into 5 main phases, that are split like this:

- **Phase 1**: Data understanding (1 week)
- **Phase 2**: Data preparation (1 week)
- **Phase 3**: Modeling (1 week)
- **Phase 4**: Evaluation (1 week)
- **Phase 5**: Deployment (1 week)